Observation errors in early historical upper-air observations

R. Wartenburger,[1] S. Brönnimann,[1] and A. Stickler[1]

Received 8 May 2013; revised 14 October 2013; accepted 16 October 2013; published 11 November 2013.

[1] Upper-air observations are a fundamental data source for global atmospheric data products, but uncertainties, particularly in the early years, are not well known. Most of the early observations, which have now been digitized, are prone to a large variety of undocumented uncertainties (errors) that need to be quantified, e.g., for their assimilation in reanalysis projects. We apply a novel approach to estimate errors in upper-air temperature, geopotential height, and wind observations from the Comprehensive Historical Upper-Air Network for the time period from 1923 to 1966. We distinguish between random errors, biases, and a term that quantifies the representativity of the observations. The method is based on a comparison of neighboring observations and is hence independent of metadata, making it applicable to a wide scope of observational data sets. The estimated mean random errors for all observations within the study period are 1.5 K for air temperature, 1.3 hPa for pressure, 3.0 ms^{-1} for wind speed, and 21.4° for wind direction. The estimates are compared to results of previous studies and analyzed with respect to their spatial and temporal variability.

Citation: Wartenburger, R., S. Brönnimann, and A. Stickler (2013), Observation errors in early historical upper-air observations, *J. Geophys. Res. Atmos.*, *118*, 12,012–12,028, doi:10.1002/2013JD020156.

1. Introduction

[2] Upper-air observations are crucial for the determination of the atmospheric state in data assimilation approaches. In order to reach an optimal analysis, it is vital to quantify the "uncertainty" (following the terminology used in data assimilation, we henceforth use the term "error") in the observations themselves ("instrument errors") and in the so-called observation operator H that represents the observations in the model space. As the errors in data assimilation systems are closely linked to the uncertainty and representativity of the observations, it is straightforward to use the observations to produce reliable estimates of both the instrument errors and the representativity errors in the assimilation (note that our definition of the term "representativity error" differs from the definition used in data assimilation; see Figure 1).

[3] A major goal of this paper is to provide error statistics that may be used in conjunction with the assimilation of historical upper-air observations. This information is important in the pre–satellite era, where upper-air observations from radiosondes, balloons, airplanes, rocketsondes, and kites are the only direct and regular measure of climate variables in the free troposphere and lower stratosphere. For early upper-air data, hardly any information on errors is available at all. Even for the past 30 years, radiosonde observation errors in present-day assimilation schemes are assumed to be merely constant in time and space for a given altitude (D. Dee, ECMWF, personal communication, 2012). Estimates of the observation error can be used to verify or complement the constant terms in the main diagonal of the error covariance matrix, which potentially improves the quality of the assimilation.

[4] A common issue for the quantification of observational errors in the pre–satellite era is the lack of suitable independent reference series. The Twentieth Century Reanalysis [*Compo et al.*, 2011] is not based on aerological observations and would hence be a potential candidate. However, the utility of this data set as a reference for the detection of errors in observational data is limited due to biases [e.g., *Stickler and Brönnimann*, 2011; *Ferguson and Villarini*, 2012]. Biases are also found within and in between different aerological data sets, even if they have been homogenized [e.g., *Xu and Powell*, 2012; *Thorne et al.*, 2011; *Francis*, 2002], highlighting that none of them can be used as a fully reliable reference.

[5] An alternative to the use of reference series for the estimation of observational errors are metadata, yet historical upper-air measurements often lack of this information. While efforts have been undertaken to compensate for some of the missing metadata [e.g., *Gaffen*, 1993], the information (which is still incomplete and partly erroneous) does not allow for a systematic estimation of observational uncertainties in historical upper-air data. Metadata also comprise observational errors estimated in previous studies. Parallel measurements are rare in the early decades and comparisons with nearby mountain sites are only possible at few locations. Based on such information, *Brönnimann et al.*

[1]Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Bern, Switzerland.

Corresponding author: R. Wartenburger, Oeschger Centre for Climate Change Research and Geographical Institute, University of Bern, Hallerstr. 12, CH-3012, Bern, Switzerland. (richard.wartenburger@giub.unibe.ch)

2169-897X/13/10.1002/2013JD020156

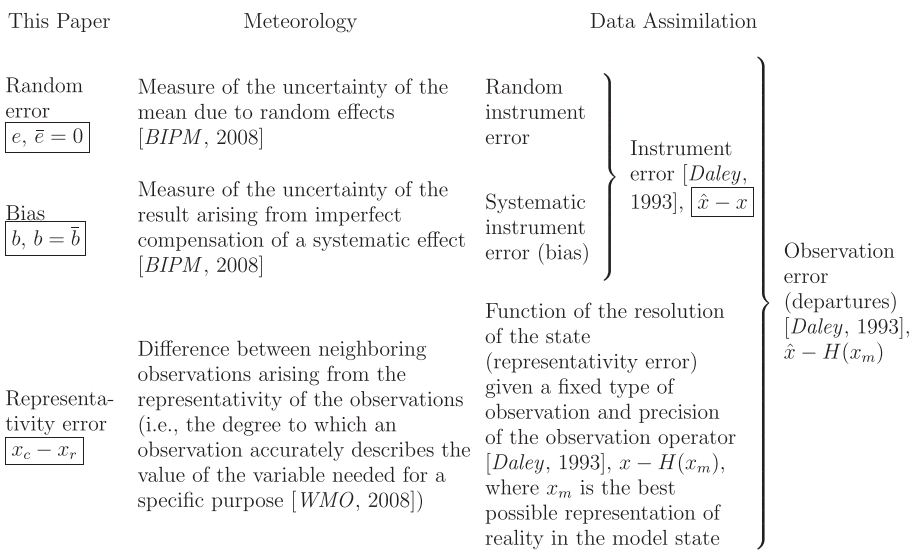

Figure 1. Comparison of terminologies for errors and uncertainties in observational data common in meteorology and data assimilation. The terms used in this paper (leftmost column) refer to observation-based estimates (middle column) of errors relevant for data assimilation (rightmost column). $\hat{x}$ represents an observation, which deviates from the unknown true state of the atmosphere x by a bias term b and a random error term e. The difference of x at two spatially distinct locations (subscripts c and r) is expressed as the representativity error of x_c with respect to x_r. The definition of the representativity error that is used in this paper differs from the one used in data assimilation, where it depends on the observation operator H, and hence on the model grid. Random errors and representativity errors are estimated from the variance, biases are estimated from the mean. A mathematical definition of the error terms is provided in section 3.2.

[2011] estimated random errors of 0.9–1.2°C and 1.35 hPa (standard deviations) for temperature and pressure measurements in German radiosonde data from the late 1930s. First comprehensive in-flight experiments for the determination of observational errors were performed in the 1950s [(*OMI*) *Organisation Météorologique Internationale*, 1951; (*OMM*) *Organisation Météorologique Mondiale*, 1952]. *Jasperson* [1982] experimentally estimated errors in wind speeds for a Doppler-based tracking system of pilot balloons. A number of studies examined sonde drift errors in greater detail [e.g., *McGrath et al.*, 2006; *Seidel et al.*, 2011], while others focused on sonde-specific radiation errors [e.g., *Brasefield*, 1948; *Rossi*, 1954]. *Kitchen* [1989] applied a comprehensive analysis of spatial and temporal representativity errors for UK Meteorological Office RS3 sondes. Although all of those studies provide valuable error statistics, they cannot be used to infer error statistics for the full range of synoptic historical upper-air observations.

[6] The issue of detecting and correcting systematic errors in upper-air data is an active topic in climate research. Most of the homogenization approaches are motivated by attempts to produce homogenized observation series that are useful for the analysis of long-term trends. A variety of techniques for break detection and adjustment were investigated. *Lanzante et al.* [2003] developed an absolute homogenization method based on a semisubjective break detection. *Free et al.* [2005] investigated the first differences technique for the reduction of inhomogeneities. *Thorne et al.* [2005] developed a relative homogenization method based on neighbor composites. This method was expanded to a fully automated homogenization by *McCarthy et al.* [2008]. Other approaches make use of innovation statistics of the European Centre for Medium Range Weather Forecast (ECWMF) 40 Year Reanalysis (ERA-40) to homogenize radiosonde temperatures [*Haimberger*, 2007; *Haimberger et al.*, 2008] and upper-air winds [*Gruber and Haimberger*, 2008]. *Sherwood* [2007] and *Sherwood et al.* [2008] address the homogenization of upper-air temperatures and wind shear of data sets that suffer from numerous gaps by applying a kriging method. Although all of these approaches succeed to produce homogenized data sets, they all differ from the goal of this paper, which is to quantify the errors of individual upper-air observations.

[7] The error estimation approach presented in this paper differs from the techniques mentioned so far. It is based on a direct comparison of observations from a candidate series to neighboring reference series using aerological data of the Comprehensive Historical Upper-Air Network (CHUAN) [*Stickler et al.*, 2010]. By these means, we avoid the shortcomings of independent reference series or metadata of questionable quality. Moreover, we are able to estimate errors more comprehensively than previously by incorporating all observations that are available in the data set.

[8] In the following section, we briefly describe the observational data that the error estimation method is applied to. Section 3 describes the error estimation method both mathematically and technically and discusses sensitivities to parameter choices. In section 4, the method is tested against a climatology and the estimated errors are discussed and compared with independent estimates. The main findings are summarized in the conclusions.

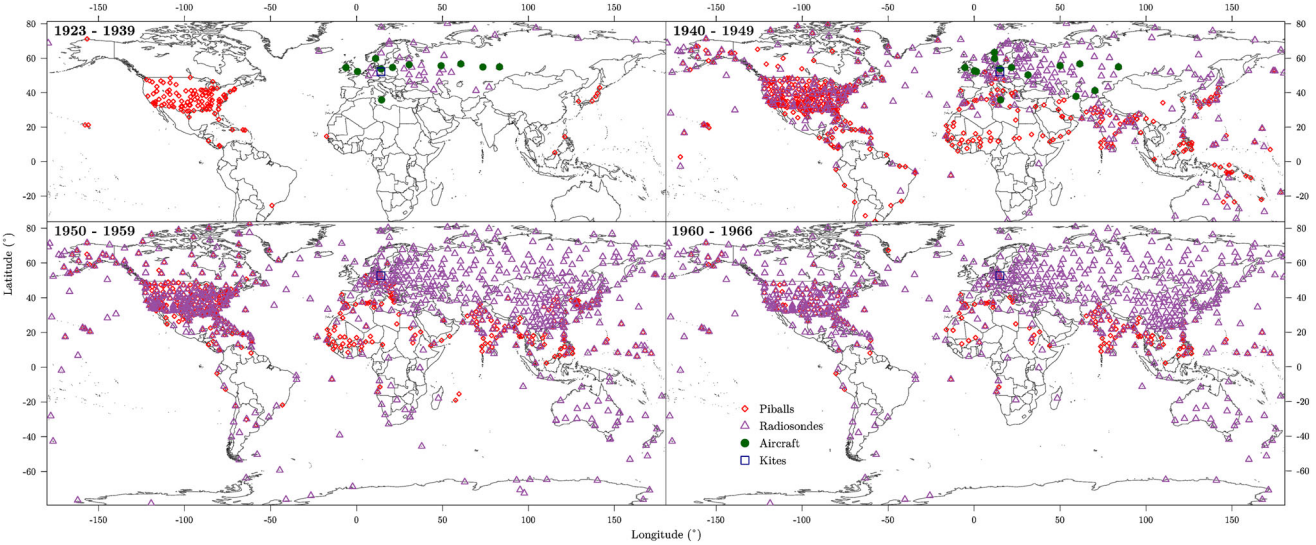

Figure 2. Locations and platform types of stations in operation during four distinct time spans as indicated on each map. Only stations with at least 30 observations are shown.

2. Observational Data

[9] We analyze errors in the CHUAN data set, version 1.70 [*Stickler et al.*, 2010]. The data set covers air temperature, geopotential height, and wind observations from 4182 stations in operation from close to the beginning of operational upper-air observations in 1904 till the end of the International Geophysical Year (IGY) in 1958 (Figure 2). Continuous records were supplemented by observations from the Integrated Global Radiosonde Archive (IGRA) [*Durre et al.*, 2006] using RAdiosone OBservation COrrection using REanalyses (RAOBCORE) 1.4 adjustments (which for the time period studied here is quasi-identical to RAOBCORE 1.5) [*Haimberger et al.*, 2012]. To include the transition time around the IGY, we analyze CHUAN observations from 1923 (the earliest year for which error estimates are available) till 1966. Each CHUAN record contains observations from either aircraft, kites, pilot balloons, registering balloons, or radiosondes. Due to different reporting practices, observations are given on pressure levels or geometric altitude levels.

[10] The adjusted version of CHUAN is used for the error analysis (see Appendix A for technical details of the adjustment). For temperature and geopotential heights, statistical break detection was performed, but adjustments were limited to breaks that were confirmed by metadata (which was rare) or coincided with a change of source (merging of different sources). Breaks were attributed to one of several predefined causes (such as a radiation error or a pressure offset) based on the shape of the vertical error profile [*Brönnimann*, 2003; *Grant et al.*, 2009]. No additional breakpoints were used in order to circumvent the lack of adequate metadata (e.g., unknown sonde types or on rare occasions even unknown platform types). For winds, the observations were compared to a reconstruction and manually screened [*Stickler et al.*, 2010, online supporting information]. Suspicious wind observations were flagged. The resulting data set is considered to be free of large biases. It is deemed to be more appropriate for a detailed estimation of errors than the raw data, as the adjusted observations resemble the characteristics of observations that would actually be assimilated (raw observations are usually quality-screened prior to assimilation).

[11] Additional tests were performed on individual observations of CHUAN to supplement the earlier quality checks, which were partly based on monthly means. The data set was checked and corrected for inconsistencies in the file format. Range checks were applied to detect physically implausible observations, leading to an adjustment of the flags of several observations. Duplicate profiles containing observations on at least three altitude levels were removed (see remark in *Grant et al.* [2009]). For the estimation of biases, we make use of anomalies from a climatology based on 6-hourly fields of the ERA-Interim reanalysis [*Dee et al.*, 2011] (technical details are provided in Appendix B).

3. Errors in Aerological Observations

3.1. Physical Error Sources

[12] Ideally, biases and random errors can be related to their physical sources. Therefore, in order to better interpret the errors, we briefly discuss the most important physical error sources. It is necessary to distinguish between the different observation platforms and between the different measurands and derived quantities (temperature T, pressure p, wind speed w, wind direction Θ, and geopotential height Z). Errors (uncertainties) in these measurands are linked to individual (partly overlapping) error sources.

[13] The magnitude of biases for both T and p is mainly controlled by the instrumentation type and by observational practices common for an individual station network. For example, stations within the same network are usually equipped with the same sonde type, whose mean radiation error differs from that common to other networks. Besides that, biases in T and p are commonly caused by a lag of the temperature sensor, by differences in the calibration, by errors in the pressure sensor, or by deposition of ice or water on the sensor [*World Meteorological Organization* (*WMO*), 2008]. For w and Θ, biases are typically related to the type

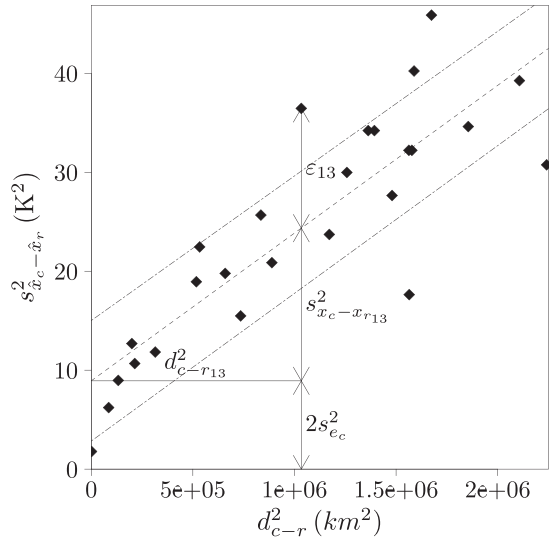

Figure 3. Schematic illustration of the assumed linear relation between $s^2_{\hat{x}_c-\hat{x}_r}$ (y axis) and d^2_{c-r} (x axis) for 500 hPa temperatures measured at the candidate station 3475 (Brunswick). The dashed line corresponds to the least squares fit, the dot-dashed lines indicate standard deviations of the residuals. The other lines illustrate the relation between the squared station separation distance $d^2_{c-r_{13}}$ (horizontal line) and the error terms used in equation (3) (variances; arrows). ε_{13} is the residual of the thirteenth data point with respect to the least squares fit, and $s^2_{x_c-x_{r_{13}}}$ is the variance of the representativity error of the candidate series for the distance $d^2_{c-r_{13}}$.

of balloon used, to the tracking device (optical theodolites or radio-theodolites), or to the assumed ascent rate of the balloon [*WMO*, 2008]. Large biases in Θ can be due to an incorrect calibration of the North direction (e.g., geomagnetic North instead of geographic North) [*Gruber and Haimberger*, 2008].

[14] Some of the above-mentioned sources for systematic errors do also cause random errors. Random errors in all variables can be caused by an imprecise assignment of observation times or by reading measured values from discrete scales [*WMO*, 2008]. For w and Θ, random errors are also introduced by an imprecise tracking of the ascending balloon and by the motion of the balloon relative to the atmosphere [*WMO*, 2008].

[15] The list of error sources presented here only has the intention to give a brief overview. Apart from the previously cited literature, a more comprehensive listing of potential error sources in upper-air observations is provided, e.g., by *Gaffen* [1994] or *Häberli* [2006].

3.2. Mathematical Error Description

[16] An error is defined as the difference between an observation $\hat{x}$ and the value of a measurand x (which is unknown) [*Bureau International des Poids et Mesures*, 2008]. It can be partitioned into a systematic part or bias b (time-invariant within the analyzed time window) and a random part e, which is assumed to be symmetrically and unimodally distributed around a zero mean

$$\hat{x} - x = b + e, \quad \text{with} \quad b = \bar{b} \quad \text{and} \quad \bar{e} = 0$$

where an overbar denotes an average over time. In many practical applications, because x is unknown, we compare the observation with a grid point value from another data set or another observation. For the difference between two observations, we can write

$$\hat{x}_c - \hat{x}_r = (x_c - x_r) + (b_c - b_r) + (e_c - e_r) \tag{1}$$

where $(x_c - x_r)$ is termed representativity error. Apart from rare exceptions (e.g., undetected copying of parts of the observations from one station series to another one), it is reasonable to assume that e_c and e_r are independent. Consequently, the variance (over time) of the difference can be written as follows:

$$s^2_{\hat{x}_c-\hat{x}_r} = s^2_{x_c-x_r} + s^2_{e_c} + s^2_{e_r} \tag{2}$$

To ensure that $s^2_{\hat{x}_c-\hat{x}_r}$ is a proper estimate for random errors, it is necessary to test if the distribution of $\hat{x}_c - \hat{x}_r$ is normally distributed (which implies symmetry and unimodality). We applied the Anderson Darling Test of Goodness of Fit [*Anderson and Darling*, 1954] and the Jarque-Bera tests [*Jarque and Bera*, 1980]. The tests suggest that 82.0 % (82.2 %) of the temperature differences, 81.6 % (81.4 %) of the geopotential height differences, 73.7 % (78.2 %) of the wind speed differences, and 88.0 % (76.0 %) of the wind direction differences are normally distributed at a level of significance of $\alpha = 10\,\%$. These results suggest that our approach is suitable for the estimation of errors in T and Z, but less for the estimation of errors in w and Θ. We still apply our approach to all variables, but advise to be cautious in the interpretation of the estimated wind errors.

[17] The focus of this study is on random errors and representativity errors. They are estimated for a given candidate observation (subscript c) from a number of neighboring (or reference) observations $(r_1, \ldots, r_n)$. We further assume, for each pair, that the random error of both observations has the same distribution (we assume that the geographical dependency of random errors from neighboring stations is negligible). Thus, we can write for the variances

$$\begin{aligned} s^2_{\hat{x}_c-\hat{x}_{r1}} &= s^2_{x_c-x_{r1}} + 2s^2_{e_c} \\ &\vdots \\ s^2_{\hat{x}_c-\hat{x}_{rn}} &= s^2_{x_c-x_{rn}} + 2s^2_{e_c} \end{aligned}$$

If we further assume that $s^2_{\hat{x}_c-\hat{x}_{r_i}}$ depends linearly on the squared Euclidean distance $d^2_{c-r_i}$ between a candidate observation and a neighboring observation i, $i = 1, \ldots, n$ (an analysis of scatterplots suggests that such a relation can indeed be postulated), a regression approach can be used to estimate $s^2_{e_c}$ and $s^2_{x_c-x_{r_i}}$

$$s^2_{\hat{x}_c-\hat{x}_{r_i}} = c_0 + c_1 \cdot d^2_{c-r_i} + \varepsilon_i \tag{3}$$

We interpret c_0 as $2s^2_{e_c}$, c_1 as $s^2_{x_c-x_{r_i}} \cdot d^2_{c-r_i}$, and ε_i as the uncertainty inherent to the model. Equation (3) can be illustrated geometrically (Figure 3). The uncertainty of the fit (which is expressed as the standard deviation of the residuals ε) is typical for candidate series with a moderate number of reference series.

[18] Apart from random errors and representativity errors, we also estimate the bias b, which we further partition into a network-wide bias b_n and a station bias b_s. For the time-average of equation (1) ($\bar{e} = 0$), we get

$$\overline{\hat{x}_c - \hat{x}_r} = \overline{(x_c - x_r)} + (b_{n_c} - b_{n_r}) + (b_{s_c} - b_{s_r})$$

Within the same network ($b_{n_c} = b_{n_r}$), we get

$$\overline{\hat{x}_c - \hat{x}_r} = (\bar{x}_c - \bar{x}_r) + (b_{s_c} - b_{s_r})$$
$$b_{s_c} - b_{s_r} = \bar{\hat{x}}_c - \bar{\hat{x}}_r - (\bar{x}_c - \bar{x}_r)$$

Considering many reference stations (which are usually located uniformly around the candidate station), we assume that b_{s_r} is statistically independent and the average over all reference series (for a particular altitude level) is zero, $[b_{s_r}] = 0$

$$[b_{s_c} - b_{s_r}] = \bar{\hat{x}}_c - \bar{x}_c - [\bar{\hat{x}}_r - \bar{x}_r]$$
$$b_{s_c} = \bar{\hat{x}}_c - \bar{x}_c - [\bar{\hat{x}}_r - \bar{x}_r]$$

For observations $\hat{x}_{n_c}$ and $\hat{x}_{n_r}$, taken at all stations within two adjacent networks, we thus get

$$\left[\overline{\hat{x}_{n_c} - \hat{x}_{n_r}}\right] = \left[\bar{x}_{n_c} - \bar{x}_{n_r}\right] + \left[b_{n_c} - b_{n_r}\right] + \left[b_{s_c} - b_{s_r}\right]$$

where we assume that the mean of all contributing individual station biases relative to the respective network biases is zero, $[b_{s_c}] = 0$ and $[b_{s_r}] = 0$

$$\left[b_{n_c} - b_{n_r}\right] = \left[\bar{\hat{x}}_{n_c} - \bar{\hat{x}}_{n_r}\right] - \left[\bar{x}_{n_c} - \bar{x}_{n_r}\right]$$

[19] The notation that is used in the rest of this paper refers to the Euclidean distance between a candidate series and an arbitrary location as d, to the random error of a candidate series as s_e with $s_e^2 = s_{e_c}^2$, and to the respective representativity error as s_{pd} with $s_{pd}^2 = s_{x_c - x_{r_i}}^2 / d_{c-r_i}^2 \cdot d^2$ (i.e., $s_{x_c - x_{r_i}}^2$ normalized by distance; $i = 1, \ldots, n$). We choose $d^2 = 10^4$ km^2 ($d = 100$ km) to approximate the mean length scales of the resolution of an ordinary model grid as used in modern atmospheric reanalyses. To preserve the original units, we show the square root of $abs(s_e^2)$ and $abs(s_{pd}^2)$. Values in single square brackets ([]) correspond to spatial averages on a single altitude level, while errors in double square brackets ([[]]) denote vertical averages.

3.3. Technical Error Description

[20] For the implementation of the theory (previous section) to real-world data (i.e., to sparse historical observations), it is necessary to define a number of threshold parameters. Obviously, the number of variance estimates $s_{\hat{x}_c - \hat{x}_{r_i}}^2$, $i = 1, \ldots, n$ is limited by the number of reference stations n in the neighborhood of a candidate station. We hence need to define an upper limit for the neighbor search radius, max(d), and a lower limit for the number of reference series, min(r_n). Individual reference series are only considered, if at least 30 of their observations overlap in time with those of the candidate series (a temporal overlap is constrained by a time window Δt centered at the observation times of $\hat{x}_c$). Multiple observation pairs within the same time window are allowed. To avoid large errors in wind directions due to low wind speeds, wind directions are only used, if the corresponding wind speeds exceed 3 ms^{-1}. For wind directions, 360° are subtracted from (added to) differences that are above (below) + (–) 180°.

[21] The determined set of optimal threshold parameters is listed in Table A1. The parameters are defined by weighing the number of candidate series c_n (which equals the number of error estimates) and the mean number of reference series [r_n] against a set of statistical measures that define the overall quality of each individual least squares model (see equation (3); details are provided in Appendix C). In parallel to the parameter selection, we tested the sensitivity of the error estimates with respect to various combinations of the threshold parameters. It was found that the error estimates are most sensitive to the choice of the neighbor search radius (max(d)), while the other threshold parameters only play a marginal role (see Appendix C). However, as the detected optimal values of max(d) are well in agreement with influence radii determined from the average 0.5 decorrelation distance (i.e., the average radius beyond which the spatial correlation drops below 0.5) for CHUAN observations in a previous study by *Griesser et al.* [2010], we can adopt the suggested values.

[22] Gross errors in the observations (e.g., large and systematic processing or digitization errors) are likely to be small in the analyzed (adjusted) version of CHUAN. However, as the quality control and homogenization procedures that were previously applied to CHUAN were partly applied on time scales greater than a month, some outliers are possibly still present in the single ascent data. For this reason, we generate a subversion of the input data set for which outliers in the observation differences $\hat{x}_c - \hat{x}_r$ are removed for all combinations of $\hat{x}_c$ and $\hat{x}_r$ for which an error estimate could be computed. As a threshold for the detection of outliers, we use the upper and lower fences f_u and f_l of the distributions of all observation differences determined per measurand, level, and distance interval (intervals range from (0, 100] km to ((max(d)–100), max(d)] km). The fences f_u and f_l are a function of the upper and lower quartiles Q_1 and Q_3: $f_u = Q_3 + k \cdot (Q_3 - Q_1)$, $f_l = Q_1 - k \cdot (Q_3 - Q_1)$ [*Frigge et al.*, 1989]. As in the case of small sample sizes, the distribution of conventional quartile estimates may not be strictly Gaussian, we use median-unbiased quartiles and the standard value of 1.5 for the factor k [*Hyndman and Fan*, 1996]. Individual differences $\hat{x}_c - \hat{x}_r$ above (below) the upper (lower) thresholds are flagged for removal. The impact of the outlier treatment on the error estimates is briefly outlined in Appendix C.

[23] Station biases b_{s_c} are determined for all candidate series with at least min(r_n) reference series. Network biases are computed for all stations with a valid network identifier and $min(r_n) = 1$ (stations using Vaisala sondes and stations with unknown network identifier were considered to belong to the Vaisala network) [*Grant et al.*, 2009]. The climatological difference $\bar{x}_c - \bar{x}_r$ is computed using a climatology of the ERA-Interim reanalysis (see Appendix B). We hereby make the assumption that this climatology is a good estimate of the mean state of the atmosphere during the study period.

4. Results

4.1. Test Against Climatology

[24] As the estimation of s_e and s_{p100} (i.e., s_{pd} for d = 100 km) is based on statistical concepts, it is useful to

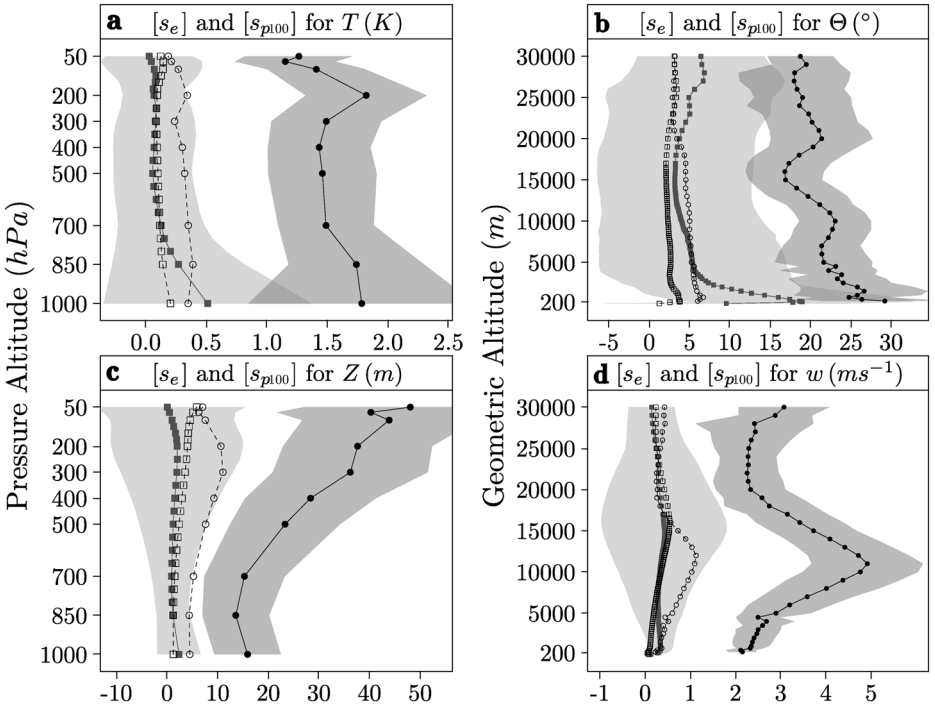

Figure 4. Profiles of mean random errors [s_e] (solid lines with filled symbols) and representativity errors [s_{p100}] (dashed lines with open symbols) for (a) temperature, (b) wind direction, (c) geopotential height, and (d) wind speed estimated from the CHUAN observations (circles) and from the ERA-Interim climatology interpolated to the locations of the CHUAN stations (squares). Shaded bands indicate the standard deviations of the random errors estimated from the climatology (light gray) and from CHUAN observations (medium gray) for all stations; their overlap is printed in dark gray. Levels with less than 30 error estimates were omitted.

validate the error estimation method. This can be achieved by applying the method to data sets with known error characteristics (the parameters are the same as for the observations, see Table A1). We make use of the interpolated and smoothed ERA-Interim climatology (Appendix B). This data set is preferred over a synthetic data set, as it contains most of the statistical properties of real observational data, but no random errors (i.e., $s_e = 0$). If our regression approach fails to reproduce $s_e = 0$, its estimates (for a given variable and altitude level) are less accurate.

[25] The error profiles are shown in Figure 4. Not surprisingly, the observation-based estimates (both [s_e] and [s_{p100}]) are substantially different from the values that were derived from the climatology. For Z, w, and T above 850 hPa, the estimated "random errors" of the climatology are approximately symmetric to zero. For wind directions, the mean of [s_e] from the ERA-Interim climatology is shifted to the right, indicating that error estimates for this variable are rather pessimistic. The amplification of [s_e] for Θ and T near the earth surface can be explained by the biased sampling of near-surface observations in mountainous areas. However, substantial deviations from zero over the entire profile (Θ) indicate that the relationship between $s^2_{\hat{x}_c-\hat{x}_{r_i}}$ and $d^2_{c-r_i}$ is not strictly linear. Combining these results with the test of $\hat{x}_c - \hat{x}_r$ for normality (section 3.2), we can conclude that the error estimates for both w and Θ are potentially biased and have to be interpreted with care. However, the test results indicate that our method does succeed to produce reliable estimates for T and Z.

4.2. Biases

[26] In the following, we discuss the station and network biases, which were estimated for the entire study period. Besides our general interest in the results, the bias estimates are useful to verify the bias adjustments that were previously applied to CHUAN (Appendix A). In this respect, the magnitude of the biases indicates both the quality of the data set and the performance of the applied adjustments.

[27] The spatial distribution of station biases of T, w, and Θ on 500 hPa (5000 m) is shown in Figure 5. Only the Northern Hemisphere is shown, as the number of estimated biases in the Southern Hemisphere is too low (the same applies to s_e and s_{pd}). Based on the design of the applied method, bias estimates in regions of high station densities are assumed to be fully reliable, while they need to be interpreted with care in regions where stations are not uniformly distributed in space (e.g., at continental margins). Temperature biases have no clear spatial patterns. Biases in winds are mostly constrained to observations within North America, while the density of overlapping observations in the other regions is mostly too low (note that the current version of CHUAN does not include a sufficient number of wind observations from the former Soviet Union). The range of station biases in wind speed and direction generally decreases with altitude, which likely indicates local differences in surface roughness that affect near-surface winds (not shown). Station biases in w (Figure 5b) over the U.S. territory are marked by a latitudinal gradient from mostly negative biases at 30°N to more positive biases at around 40°N. Similar features can

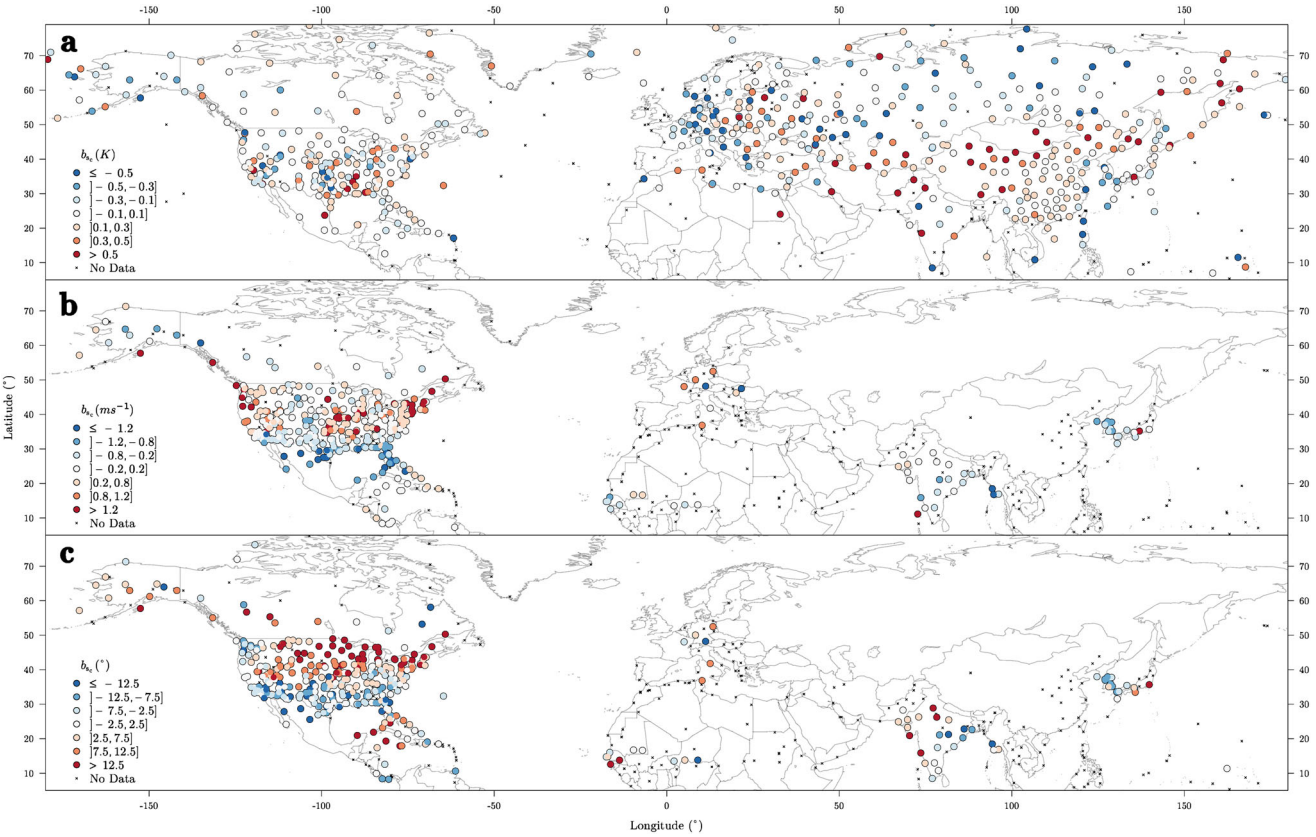

Figure 5. Spatial distribution of station biases (colored dots) for (a) 500 hPa temperatures, (b) 5000 m wind speeds, and (c) 5000 m wind directions. Small black crosses denote the location of stations for which errors were not determinable.

be found for Θ, where we find a prominent cluster of positive biases at around 45°N. The more isolated biases in Θ (which are constant throughout the entire profile) indicate systematic differences that might arise from the choice of the wrong North direction. While these isolated biases are clear indicators of systematic instrument or processing errors in individual observation series, the spatially more coherent biases could also be related to biases in the ERA-Interim reanalysis or to long-term changes in the patterns of the observed atmospheric fields. The confidence in these interpretations could certainly be increased by considering additional data sets as a reference, which is certainly an interesting perspective for future work.

[28] Figure A4 shows the profiles of network temperature biases for all combinations of neighboring station networks. The biases are generally low, but clearly depend on the choice of the networks. The spread (±1 standard deviation) among the differences of each observation pair is substantial, which underlines that the biases may be of different magnitude when considering shorter time spans or regional

Table 1. List of Mean Estimates (Variances), Mean Standard Errors of the Estimates (SE), Mean Coefficients of Determination (R^2, unitless), Mean Errors and Their Mean Horizontal Standard Deviations (sd) for 500 hPa (T, Z, p) and 5000 m (w, Θ)[a]

		T(K^2)	Z(m^2)	w((ms^{-1})^2)	Θ((deg)^2)
Estimates	$[s_e^2]$	4.65 (2.76)	1299 (855)	17.8 (11.9)	1004 (189)
	$[s_{p100}^2]$	0.0011 (0.0008)	0.63 (0.46)	0.0039 (0.0032)	0.29 (0.36)
Standard errors	$[\mathrm{SE}(s_e^2)]$	1.12 (0.66)	555 (351)	3.19 (1.75)	225 (157)
	$[\mathrm{SE}(s_{p100}^2)]$	0.0002 (0.0001)	0.090 (0.056)	0.0009 (0.0005)	0.051 (0.035)
Model fit	$[R^2]$	0.65 (0.74)	0.68 (0.72)	0.48 (0.65)	0.64 (0.84)
		T(K)	p(hPa)	w(ms^{-1})	Θ (deg)
Mean errors	$[s_e]$	1.46 (1.14)	1.58 (1.33)	2.90 (2.40)	21.64 (8.59)
	$[s_{p100}]$	0.32 (0.28)	0.52 (0.44)	0.60 (0.56)	5.24 (5.91)
Horizontal sd	$[\mathrm{sd}(s_e)]$	0.45 (0.28)	0.74 (0.47)	0.77 (0.45)	6.05 (6.17)
of mean errors	$[\mathrm{sd}(s_{p100})]$	0.08 (0.06)	0.16 (0.12)	0.18 (0.13)	1.40 (1.12)

[a]Values outside parentheses correspond to RAW and NF, values in parentheses correspond to RAW+OC and NF+OC.

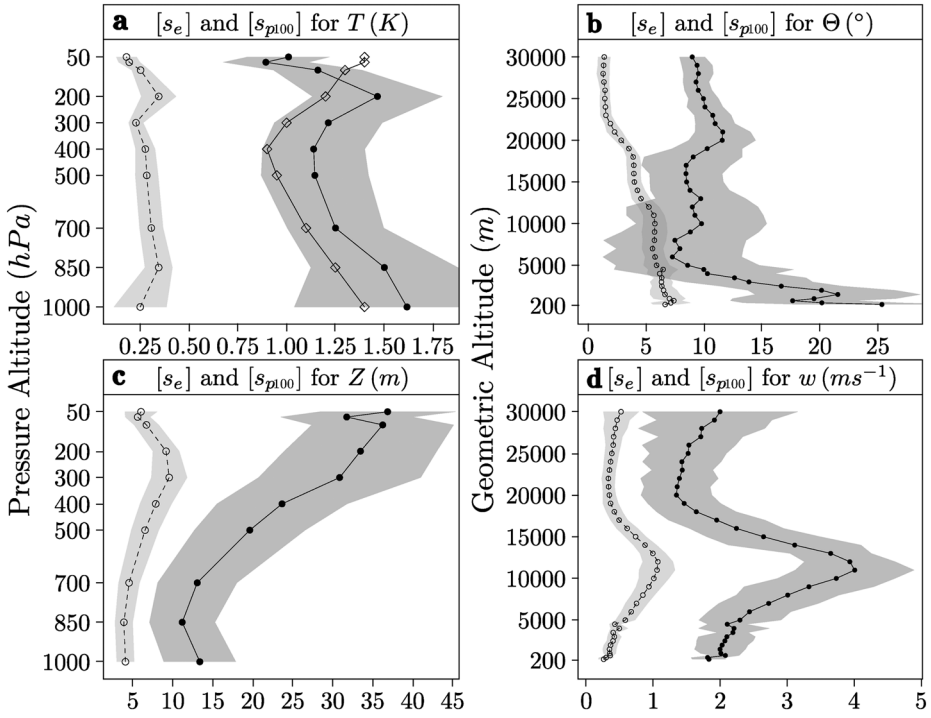

Figure 6. Profiles of mean random errors [s_e] (solid lines with filled circles) and representativity errors [s_{p100}] (dashed lines with open circles) for (a) temperature, (b) wind direction, (c) geopotential height, and (d) wind speed of the OC version. Open diamonds in Figure 6a correspond to observation errors assumed in the ERA-Interim reanalysis (see text). Shaded bands indicate the standard deviations of the random errors (medium gray) and representativity errors (light gray) for all stations; their overlap is printed in dark gray. Levels with less than 30 error estimates were omitted.

subsets. Most of the temperature biases are in the range of ±1 K from the mean (and hence within the range of the station biases). However, for some combinations of networks, the biases clearly exceed the magnitude of the station biases. This indicates that, despite the previous adjustments applied to CHUAN, systematic differences between the station networks still exist in the data set. Biases on the 70 hPa and 50 hPa levels are particularly large, which either indicates persistent radiation errors in the Vaisala and U.S. networks, or an overcorrection of radiation errors in observations from the Soviet networks.

4.3. Random Errors and Representativity Errors

[29] This section presents and discusses the final version of the error estimates (T and Z: all available observations (RAW), w and Θ: all except flagged observations (NF); see Appendix C). In cases where the focus is on the mean error patterns, we also present the outlier-corrected (OC) estimates. In the following, we first discuss the mean errors and then analyze their spatial and temporal variability. If not denoted explicitly, the presented estimates always refer to the entire study period.

4.3.1. Mean Errors

[30] Table 1 lists the estimated variances, the corresponding error estimates, their standard deviations (i.e., the spread of all error estimates), and parameters that describe the uncertainty of the error estimates for the 500 hPa (5000 m) level. These levels were chosen, as they are often used to represent the large-scale dynamic state of the atmosphere and are neither influenced by the planetary boundary layer nor by tropospheric jets. The pressure errors are derived by conversion of the geopotential height errors using the 1976 U.S. Standard Atmosphere [*NOAA et al.*, 1976]. Both the standard error of the estimates and the coefficient of determination indicate that the error estimates of the OC version are less noisy than the RAW (NF) estimates. In addition, [s^2_{p100}] is less sensitive to the outlier removal than [s^2_e], which underlines that the OC version can be used to analyze mean error profiles. For all other analyses, we use RAW (NF), as it captures the full magnitude of [sd(s_e)] and [sd(s_{p100})].

[31] The mean profiles of the random and representativity errors of the OC data allow for the identification of factors that dominate the errors throughout time and space (Figure 6). For comparison, we also plotted the observation errors for temperatures that are used in the ERA-Interim reanalysis. This error is assumed to be constant for all assimilated radiosonde observations on a given pressure level, whereas no distinction is made for different sonde types or instrument designs (P. Poli, ECMWF, personal communication, 2012). The representativity error of the data assimilation system (which is a component of the observation error; cf. Figure 1) is arguably smaller than the estimated representativity errors, as the horizontal resolution of ERA-Interim (T255) corresponds to grid cell distances that are (in average) smaller than the separation distance d = 100 km. Throughout the troposphere, the estimated random errors are larger than the observation errors, while their vertical structure is well resembled. Given that the observation errors were specified at the time the ERA-Interim reanalysis was built, this result is in agreement

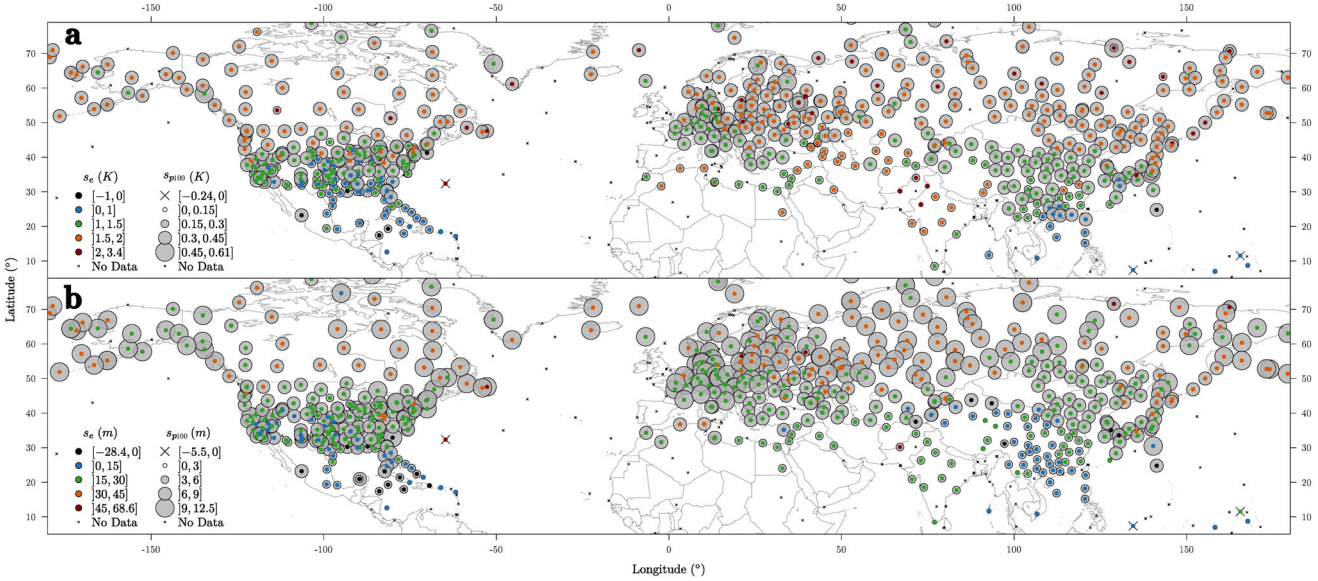

Figure 7. Spatial distribution of random errors (colored dots) and representativity errors (gray circles) for (a) 500 hPa T and (b) 500 hPa Z. Small black crosses denote the location of stations for which errors were not determinable.

with the long-term increase in the accuracy of observations over time (cf. Figure 10). On altitudes above 200 hPa, we find a substantial disagreement of the errors that requires further attention.

[32] The mean temperature representativity errors [s_{p100}] (Figure 6a) show maxima on the 850 hPa and 200 hPa levels. The 850 hPa peak is located within the global average height of the planetary boundary layer (500–2000 m, according to *Seidel et al.* [2010]), while the 200 hPa peak falls within the annual mean height of the midlatitude tropopause (≈ 150–250 hPa, according to *Hoinka* [1998]). Both peaks are most likely caused by height variations in these atmospheric features. In accordance to the high degree of spatial homogeneity of the temperature fields in the upper troposphere and lower stratosphere, representativity errors at these altitudes are low. The mean random errors are largest near ground and show no peak on 850 hPa, which indicate that they are mostly dependent on the observing system. They are, however, not fully independent of ambient weather conditions. For instance, wrongly corrected instrument lags may lead to lower errors in areas where lapse rates are rather constant such as in the middle and upper troposphere. The spread in [s_e] indicates the heterogeneity of the contributing observations (note that the spread is larger in the RAW version).

[33] The mean profile of [s_{p100}] for Θ (Figure 6b) is mainly characterized by a distinct decrease with height up to an altitude of ≈ 23 km interrupted by a secondary maximum at an altitude of around 9–11 km (≈ 307–226 hPa), which is just below the annual mean height of the midlatitude tropopause. Despite the limited validity of the error estimation method for use with wind directions, the large magnitude of the vertical variation of [s_{p100}] suggests a predominance of natural, i.e., climatological factors. Random errors of Θ are higher near ground, where wind directions are more heterogeneous due to different station altitudes and due to the influence of surface roughness. The standard deviation over all stations is low above 20 km and largest in the middle and upper troposphere.

[34] For geopotential height (Figure 6c), the strong increase of the tropospheric representativity errors with altitude is dominated by an increase of the mean heights with altitude, which cause an increase in spatial differences of Z measured within the radius max(d). Within the lower stratosphere, the geopotential height gradients decrease again. The increase of [s_e] with altitude is due to the summation of the errors in the pressure and temperature measurements during the radiosonde ascents. The lower value on the 70 hPa level is related to the fact that this pressure level was mainly reported in radiosonde ascents from the end of the study period (which are characterized by smaller random errors). The spread of [s_e] is low compared to the heterogeneous distribution of the respective representativity errors.

[35] The shape of the error profiles of w (Figure 6d) is dominated by the average altitude of the maximum wind speeds on ≈ 12 km. The increase of [s_e] up to this altitude can also be explained by the dependency of the angular errors on the total distance between the theodolite and the balloon. In the stratosphere, however, wind speeds are generally less strong and more homogeneous, causing a substantial decrease of both random errors and representativity errors. The magnitude of the spread is correlated to the magnitude of the mean errors.

4.3.2. Spatiotemporal Error Structure

[36] The analysis of spatial error patterns is assisted by the use of error maps. Figures 7 and 8 show the individual random errors and representativity errors of all analyzed variables on the 500 hPa (5000 m) level. Errors on this altitude are not substantially different from the neighboring levels and are therefore considered to be characteristic for the midtroposphere. What stands out for T and Z is a mean increase of the representativity errors from the subtropics to the subpolar region (see also Figure 9). This feature is in line with atmospheric circulation patterns that determine

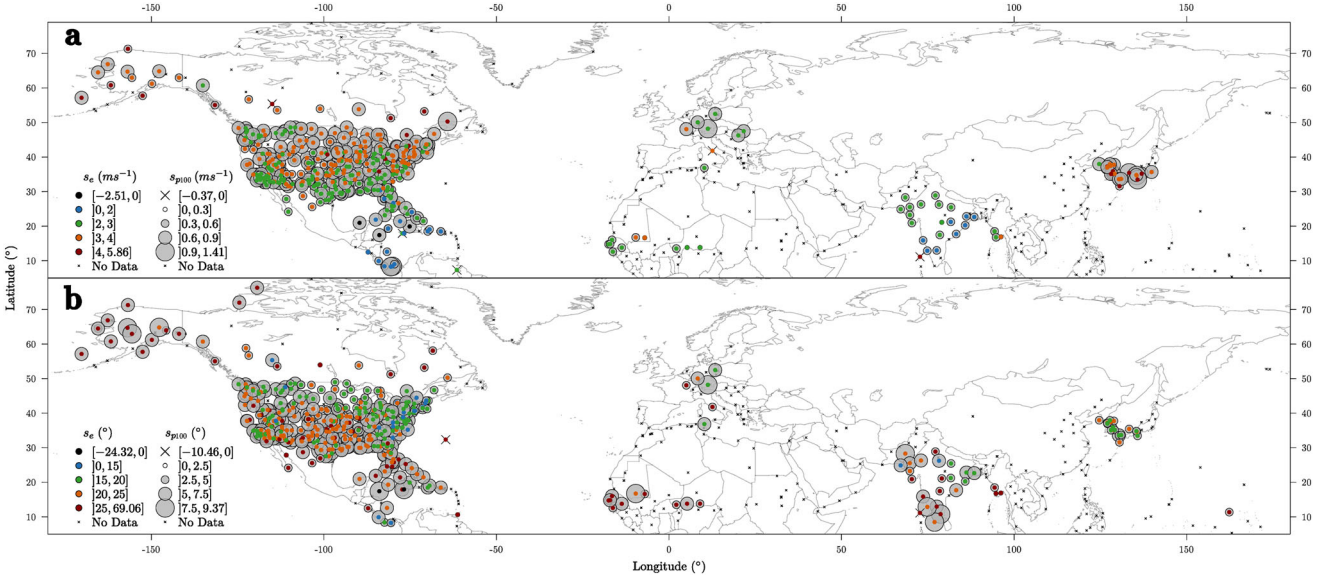

Figure 8. Spatial distribution of random errors (colored dots) and representativity errors (gray circles) for (a) wind speed and (b) wind direction on 5000 m. There are no estimates for the former Soviet Union, as radiosonde winds in CHUAN V1.7 are too sparse, and the Soviet network in CHUAN V1.7 does not contain pilot balloons. Small black crosses denote the location of stations for which errors were not determinable.

the degree of spatiotemporal variability of the atmospheric fields (e.g., the position of the jet streams). No clear latitudinal trend in $[s_{p100}]$ is found for w and Θ. Random errors of T and Z are dominated by above-average values over large parts of the former Soviet Union, while they are generally low over China, south-western Europe, the USA, and the Caribbean. This pattern is in agreement with the assumptions of the method (s_e is the same for any reference series of a given candidate series), i.e., random errors are mostly the same for neighboring stations, and differences mainly occur on longer spatial scales. This does also correspond to the expectation that random errors mainly differ in between different station networks. For winds, a large number of non-U.S. stations cover only short and nonoverlapping time spans, impeding a global comparison. Random errors in w (Figure 8a) tend to be very low in the (sub) tropical calm zones and increase toward the North. This is in contrast to random errors in Θ (Figure 8b), which are spatially more heterogeneous and large in both the subtropical and the subpolar regions. Considering the limited performance of the error model for winds, these patterns should, however, not be overinterpreted.

[37] In order to quantify the dependency of the error estimates on the station latitude, all estimates were zonally averaged over 10 degree bins (see Figure 9 for errors in geopotential heights). Both $[s_e]$ and $[s_{p100}]$ tend to increase with (Northern Hemisphere) latitude. This is linked to the location of the dominating pressure cells and associated flow features. The meridional gradient of atmospheric variability obviously explains the gradients of both errors, as an

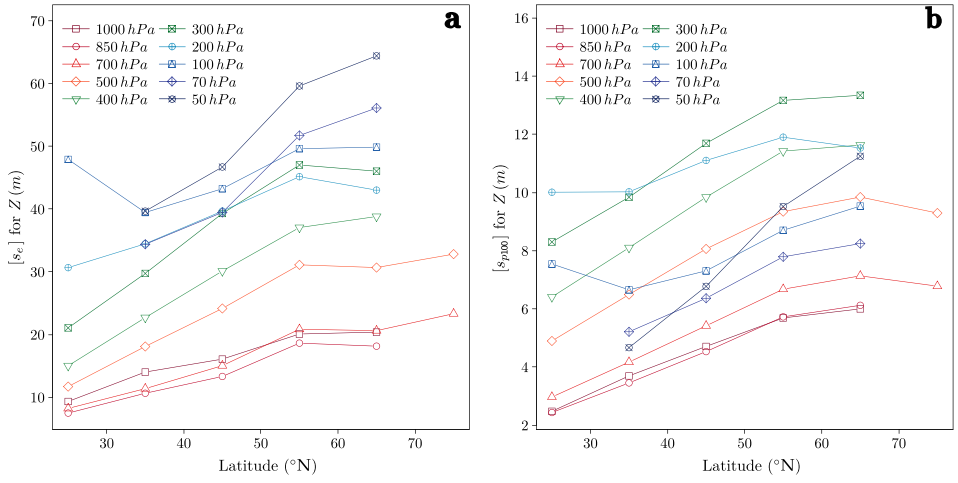

Figure 9. Zonal 10 degree mean (a) random errors and (b) representativity errors of geopotential height on selected pressure levels. Errors are plotted in the middle of the 10 degree bands.

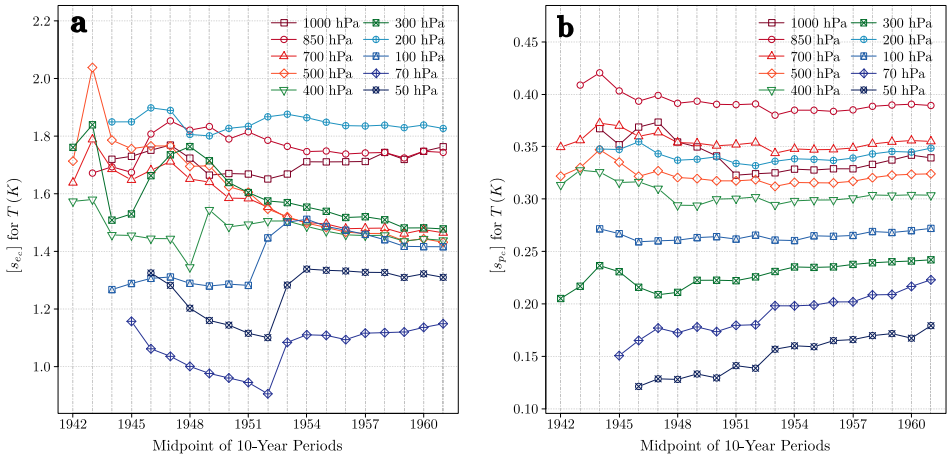

Figure 10. Time series of running decadal mean (a) random errors and (b) representativity errors of temperatures on selected pressure levels. Errors based on less than 30 error estimates were omitted. The years on the abscissa correspond to the middle of each time window.

increase in atmospheric variability leads to a mean increase in the spatial differences of the observed heights, while it amplifies the risk of random errors. The zonal mean pattern of the other variables is similar to the one of Z, though the latitudinal differences are less pronounced (not shown).

[38] The errors that were presented so far are estimated for observations from the entire study period. A deeper insight into the temporal evolution of the errors is gained by estimating the errors over shorter time spans. We choose moving 10 year windows to estimate the errors, although it would certainly be beneficial to select those periods according to the dates of potential breakpoints, which could be provided by concise metadata. Figure 10 demonstrates the evolution of $[s_e]$ and $[s_{p100}]$ for temperatures on selected pressure levels. Due to the extremely low station density before the 1940s, no error estimates are available from decades before 1938–1947. While $[s_e]$ is almost time independent near the earth surface and on the average altitude of the midlatitude tropopause (200 hPa), it decreases with time on most of the other levels. In contrast, $[s_{p100}]$ is mostly constant over all subperiods and all levels, which supports the correctness of the estimates. The substantial increase of $[s_e]$ in the early 1950s in the stratosphere can only be explained by a sudden increase in the number of radiosonde ascents that reach levels above 300 hPa, indicating that the estimated s_e for earlier periods are too small. Whether the variations in the 1940s are an actual result of changing instruments or measurement procedures could only be answered, if more comprehensive and reliable metadata was available.

[39] We also analyzed the effect of the intra-annual sampling of observations on the error estimates. Figure 11 shows the seasonal variation of $[s_e]$ and $[s_{p100}]$ for temperatures together with the corresponding standard deviations. While the representativity errors are characterized by a strong seasonal cycle, the random errors are less affected. In the troposphere, the largest random errors occur in NH winter (DJF), while the smallest errors occur in NH summer (JJA). This result is inverse to the seasonal pattern of biases (which contain radiation errors that are highest during NH summer) and can be explained by the dependency of random errors on the intra-annual cycle of atmospheric flow features and related weather conditions. Consequently, random errors above the 300 hPa level are mostly insensitive to the season.

4.4. Comparison to Other Studies

[40] The mean estimates of all errors can (to some extent) be compared to independent error estimates for historical aerological observations from previous studies. As most of these studies follow a less comprehensive approach than our method, differences to our results may in part be due to sampling errors or to methodological differences. Table 2 provides a direct comparison of the estimated errors.

[41] Temperature and pressure biases in *OMM* [1952] were estimated from in-flight comparisons that were performed in May 1950 using radiosonde types that were in

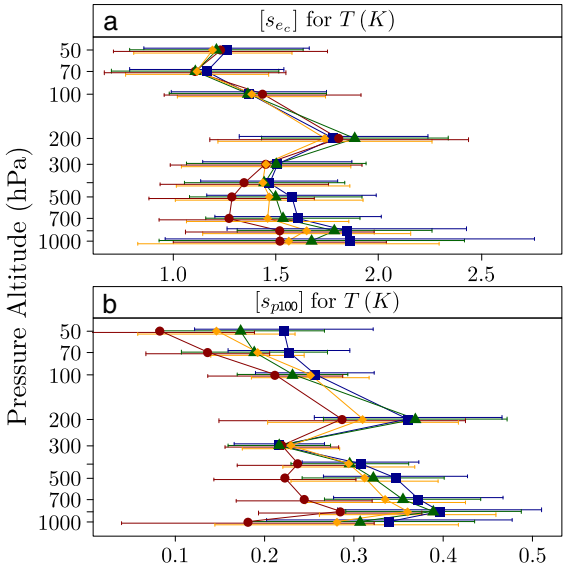

Figure 11. Profiles of seasonal mean of (a) random errors and (b) representativity errors and corresponding standard deviations (vertically displaced error bars) for DJF (blue squares), MAM (green triangles), JJA (red circles), and SON (orange diamonds) temperatures.

Table 2. Error Estimates for (Historical) Aerological Observations for Given Pressure Levels (in hPa)[a]

Source	Error Type and Variable	Vertical Mean	> 300	< 300	700	500	300	200	100	50	900–700	700–500	500–300	300–200
RAW	$[b_{n_{FI}} - b_{n_r}]$ T(K)	0.13									0.18	0.23	0.15	0.17
OMM	$[b_{n_{FI}} - b_{n_r}]$ T(K)	0.79									0.60	0.60	0.75	1.05
RAW	$[b_{n_{U.S.}} - b_{n_r}]$ T(K)	–0.22									0.10	–0.18	–0.50	–0.46
OMM	$[b_{n_{U.S.}} - b_{n_r}]$ T(K)	–0.28									0.03	–0.10	–0.43	–0.58
RAW	$[b_{n_{FI}} - b_{n_r}]$ p(hPa)	0.49									0.59	0.70	0.63	0.41
OMM	$[b_{n_{FI}} - b_{n_r}]$ p(hPa)	4.4									1.5	1.5	5.5	6.5
RAW	$[b_{n_{U.S.}} - b_{n_r}]$ p(hPa)	–0.04									0.48	0.44	0.11	–0.29
OMM	$[b_{n_{U.S.}} - b_{n_r}]$ p(hPa)	–0.56									–2.30	–1.00	–0.30	0.80
RAW	$[s_e]$ T(K)	1.50	1.58	1.41	1.49	1.46	1.49		1.41	1.27				
Brönn.	$[s_e]$ T(K)	0.9–1.2												
OMM	$[s_e]$ T(K)		0.77	1.25										
Zait.	$[s_e]$[b] T(K)				0.50	0.70	0.90		0.65	0.40				
RAW	$[s_e]$ p(hPa)	1.26	1.57	0.77										
Brönn.	$[s_e]$ p(hPa)	1.35												
OMM	$[s_e]$ p(hPa)		5.75	8.33										
RAW	$[s_e]$ $Z(m)$				15.3	23.3	36.2		43.9	48.1				
Zait.	$[s_e]$[b] $Z(m)$				5.0	9.5	26.0		38.0	48.0				
RAW	$[s_{p220}]$ T(K)				0.8	0.7	0.5	0.8	0.6	0.4				
Kit.	$[s_{p220}]$[b] T(K)				11.5	16.7	24.3	23.4	16.6	15.5				
RAW	$[s_{p220}]$ Z(m)				11.5	16.7	24.3	23.4	16.6	15.5				
Kit.	$[s_{p220}]$[b] Z(m)				26	37	55	50	40	38				

[a]The source of the data is indicated in the first column (OMM: [OMM (Organisation Météorologique Mondiale), 1952], Brönn.: *Brönnimann et al.* [2011], Zait.: *Zaitseva* [1993], Kit.: *Kitchen* [1989]). The subscript FI used for the network biases indicates the Finnish (Vaisala) network, the subscript U.S. indicates the U.S. network, the subscript *r* indicates all other networks.
[b]Error estimates are root-mean-square errors.

operation at that time. The biases were calculated as mean differences from one sonde type with respect to all other sonde types for daytime and nighttime ascents below and above the 300 hPa level. Biases in T for the U.S. sonde type are very close to our estimates, while the magnitudes of the other biases are clearly larger, possibly owing to the adjustments applied in CHUAN.

[42] The random errors of *Brönnimann et al.* [2011] are within the range of our estimates, while there is considerable disagreement to the estimates of *OMM* [1952] (in particular for pressure). The large magnitude of the pressure error estimates in the OMM intercomparison could be related to unit errors. *Zaitseva* [1993] estimated random errors for temperatures observed by historical sonde types used in the former Soviet Union, which are considerably lower than our mean estimates. This difference is even more evident if we only consider stations located within the former Soviet Union (cf. Figure 7).

[43] The representativity errors estimated by *Kitchen* [1989] for temperatures and geopotential heights are substantially higher than our estimates of $[s_{p220}]$, while their vertical structure is the same. The disagreement in the error magnitudes can be explained by methodological differences (use of root mean squared differences instead of variances) and by the smaller number of soundings that the estimates of *Kitchen* [1989] are based on.

5. Conclusions

[44] We developed an algorithm to systematically estimate random errors and representativity errors of historical upper-air wind, air temperature, and geopotential height (pressure) observations which we applied to observations from the Comprehensive Historical Upper-Air Network (CHUAN) from 1904 till 1966 (earliest error estimates possible from 1923). The error estimation method is based on a comparison of neighbor series which neither requires comprehensive metadata nor other independent data sets, but complements and confirms metadata-based analyses. It can readily be applied to other four-dimensional atmospheric observational data sets.

[45] The estimated error magnitudes are in good agreement with some studies [e.g., *Brönnimann et al.*, 2011], but not others (which mainly estimated errors of lower magnitude). The spatiotemporal dependence of the errors agrees with theoretical expectations and also shows new features. Biases between station networks generally exceed biases between neighboring station series, indicating the presence of systematic differences between the networks. The estimated representativity errors show substantial latitudinal and seasonal variations for most of the variables. Their mean profile is in agreement with the mean vertical structure of the atmosphere, while random errors are much less dependent on the atmospheric state. The random errors of the historical observations are mostly larger than observation errors assumed in a modern reanalysis product.

[46] Both the error estimation method and the estimated error statistics can be used for data assimilation approaches. We provide information about the spatial and temporal relations of random errors and representativity errors, which is important when attempting to assimilate historical upper-air data. In additional analyses, spatial and temporal error covariances could be computed from the estimated errors, which are assumed to be zero in current assimilation schemes.

[47] Apart from its utility for the reanalysis community, the results can be applied to estimate errors in existing data sets by providing uncertainty measures for (historical) data sets that do not yet contain such information. Moreover, the error estimates are suitable for a qualitative comparison to

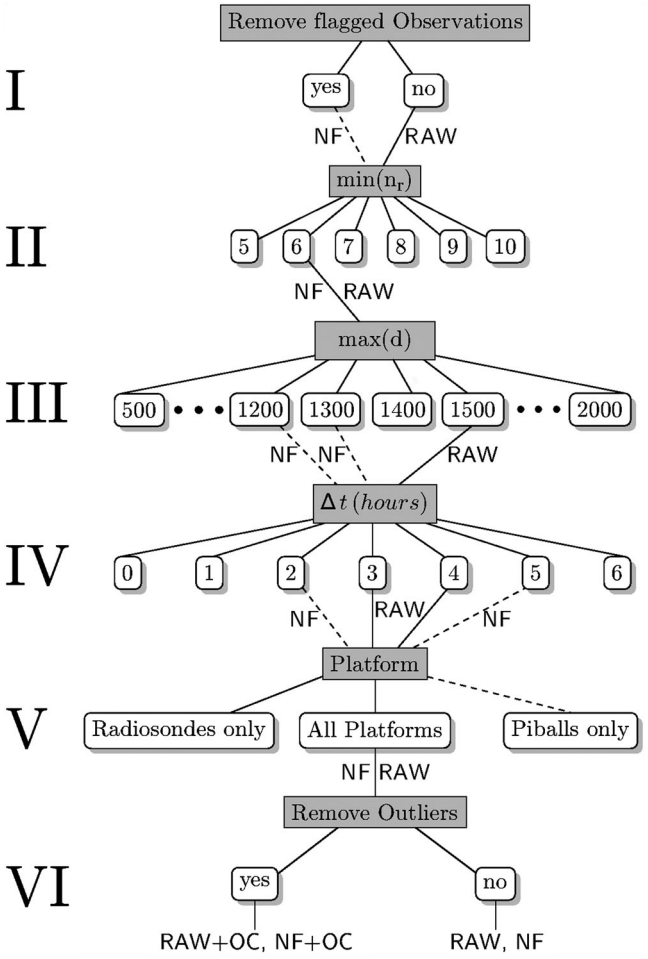

Figure A1. Partial dendrogram of the individual decisions I to VI (connecting lines) made for all parameters (gray boxes) and individual values that were tested (white boxes). Displayed are only the branches that were used for further analysis. The abbreviations in capital letters correspond to individual versions of the data sets: RAW (all available CHUAN observations), NF (CHUAN observations excluding flagged observations), and OC (CHUAN observations excluding outliers). Dashed lines correspond to decisions related to wind observations only.

error estimates of reanalysis products (e.g., the ensemble spread or analysis departures of assimilated observations). Certainly, it would be rewarding for more expansive data validation efforts to develop a gridded data product from the current point data (comparable to, e.g., Hadley Centre Atmospheric Temperature Data Set Version 2 (HadAT2); *Thorne et al.*, 2005).

[48] The method that was presented in this paper is designed to be applied to other compilations of upper-air data. A very promising candidate is the IGRA data set, which contains more extensive metadata that may aid to examine random errors and biases in greater detail. Another prospective data sets are new compilations of historical upper-air observations from the ongoing ERA-CLIM project. This data set (which is planned to be included in the next version of CHUAN (version 2.0)) will possibly allow for the estimation and analysis of observation and representativity errors in regions that are currently blank.

Appendix A: CHUAN Adjustments

[49] The adjusted versions of the CHUAN data set that are used in this paper (C.DC for wind, C.DCR for temperature, and geopotential height) were generated from the raw data by applying a number of preprocessing steps. Suspicious observations of geopotential height and temperature observations were flagged according to their quality [*Stickler et al.*, 2010, online supporting information]. Monthly means of air temperature, geopotential height, wind direction, and wind speed were reconstructed using a variety of independent predictors [*Brönnimann*, 2003]. In order to test the geopotential height and temperature observations for the presence of artificial breakpoints, each of the station series was compared to reconstructed fields [*Grant et al.*, 2009; *Griesser et al.*, 2010]. Depending on the magnitude of the errors, as well as on the skill of the statistical models, the station records were either (entirely or partly) rejected, adjusted, or accepted without adjustments. In addition to statistical tests based on the comparison with a reconstruction

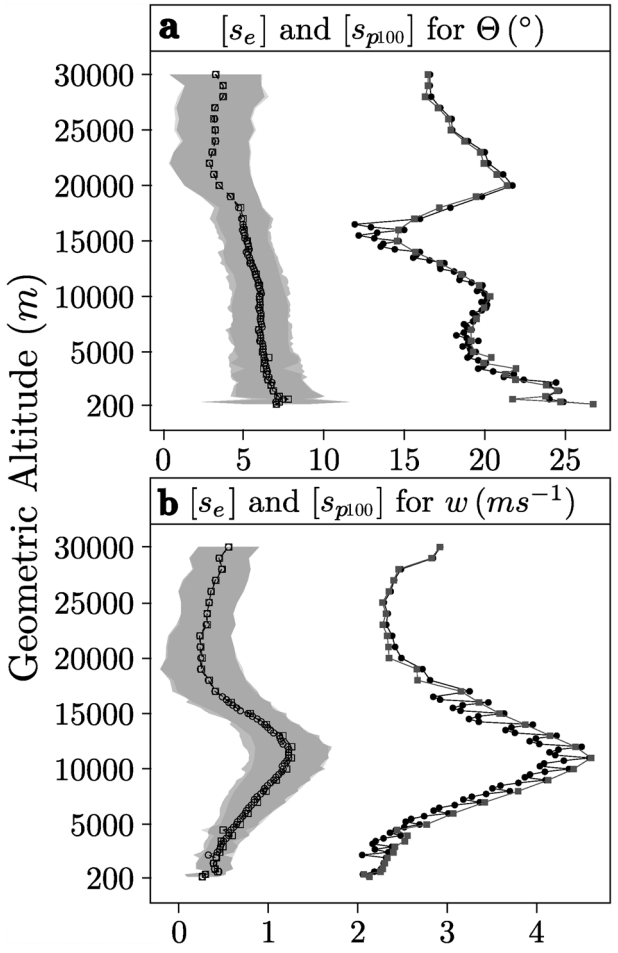

Figure A2. Profiles of mean random errors [s_e] (solid lines with filled symbols) and representativity errors [s_{p100}] (dashed lines with open symbols) for (a) wind direction and (b) wind speed of RAW (circles) and NF (squares). Shaded bands indicate the standard deviations of the random errors estimated from RAW (medium gray) and NF (light gray) for all stations; their overlap is printed in dark gray. Levels with less than 30 error estimates were omitted.

Table A1. Optimal Values for the Threshold Parameters Used in the Error Estimation Method (Minimum Number of Reference Series min(r_n), Maximum Separation Distance max(d), and Time Window Used to Treat Observations as Simultaneous Δt) Determined by the Total Number of Error Estimates c_n, the Mean Number of Reference Series [r_n], the Mean Coefficient of Determination [[R^2]], the Mean Coefficient of Variation [[cv]], and the Mean p Value of the F Test of Goodness of Fit [[p_F]][a]

Variable	Parameter	c_n ↑	[r_n] ↑	[[R^2]] ↑	[[cv]] ↓	[[p_F]] ↓	Decision
T	min(r_n)	**4**	10	7	4	10	6
Z	min(r_n)	**4**	10	6	4	10	6
w	min(r_n)	**4**	10	9	4	10	6
Θ	min(r_n)	**4**	10	10	4	10	6
T	max(d) (km)	*2000*	*2000*	1500	1000	2000	1500
Z	max(d) (km)	*2000*	*2000*	1600	1000	2000	1500
w	max(d) (km)	*2000*	*2000*	1100	700	1800	1200
Θ	max(d) (km)	*2000*	*2000*	1300	800	1900	1300
T	Δt (h)	6	6	0	1	0	3
Z	Δt (h)	6	6	0	6	0	4
w	Δt (h)	4	6	2	6	6	5
Θ	Δt (h)	1	1	1	6	1	2

[a]The range of tested values is indicated in Figure A1. Arrows indicate whether the threshold parameters were selected for the lowest (arrow pointing downward) or highest (arrow pointing upward) values that were tested. The final selection of a threshold parameter (rightmost column) corresponds to the average of columns 3–7 rounded to the next tested value. Values printed in italics were not considered, while values printed in bold were weighted times 4.

similar to *Griesser et al.* [2010], wind observations were also evaluated by means of a detailed visual inspection [*Stickler et al.*, 2010]. The applied adjustments for air temperature and geopotential height compensate for radiation and lag errors (i.e., an error due to the lagged response time of the sensor), erroneous units, pressure errors, as well as for constant temperature offsets [*Grant et al.*, 2009; *Brönnimann*, 2003]. If possible, physics-based adjustments were applied to the radiosonde observations to account for sonde-specific error characteristics [*Stickler et al.*, 2010].

Appendix B: ERA-Interim Climatology

[50] We used 6-hourly ERA-Interim fields to generate a climatology of the same spatiotemporal resolution that is based on the reference period 1981–2010. The u and v wind fields were converted to wind speed and direction and linearly interpolated to the height of the geometric altitude levels used in CHUAN by using the gravity-corrected geopotential height fields as a reference. After this step, all fields were spatially (bilinearly) interpolated to the geographical locations of the CHUAN stations. Then, the interpolated climatologies for each of the main synoptic hours (00, 06, 12, and 18 UTC) were filtered using a circular 31 day running mean filter with equal weighting. For each observation series, the smoothed climatologies were then interpolated to the respective ascent times by using natural splines to simulate the daily cycle. Anomalies were generated by subtracting the difference between the observations and the climatology.

Table A2. Range of [s_e] and [s_{p100}] for All Variables on Selected Pressure (Geometric Altitude) Levels Under Variation of the Threshold Parameters min(r_n), max(d), and Δt

		min(r_n) ∈ [4, . . . , 10]		max(d) ∈ [500, . . . , 2000] km		Δt ∈ [0, . . . , 6] h	
		[s_e]	[s_{p100}]	[s_e]	[s_{p100}]	[s_e]	[s_{p100}]
T_{850}	(K)	[1.43, 1.46]	[0.48, 0.49]	[1.18, 2.08]	[0.32, 0.66]	[1.77, 1.78]	[0.39, 0.39]
T_{500}	(K)	[1.18, 1.25]	[0.38, 0.39]	[0.93, 1.75]	[0.26, 0.47]	[1.46, 1.48]	[0.32, 0.32]
T_{300}	(K)	[1.25, 1.34]	[0.30, 0.31]	[0.96, 1.68]	[0.19, 0.40]	[1.49, 1.51]	[0.24, 0.24]
T_{100}	(K)	[1.28, 1.32]	[0.30, 0.31]	[1.10, 1.58]	[0.24, 0.38]	[1.41, 1.42]	[0.27, 0.28]
T_{50}	(K)	[1.12, 1.24]	[0.18, 0.23]	[1.00, 1.38]	[0.18, 0.24]	[1.25, 1.29]	[0.20, 0.20]
Z_{850}	(m)	[10.4, 11.1]	[4.8, 0.50]	[8.7, 17.9]	[3.8, 5.3]	[13.7, 14.2]	[4.4, 4.4]
Z_{500}	(m)	[17.9, 18.9]	[8.4, 8.8]	[15.4, 31.3]	[6.6, 9.3]	[23.9, 24.6]	[7.7, 7.7]
Z_{300}	(m)	[28.8, 29.8]	[12.5, 12.8]	[23.2, 48.0]	[9.5, 14.1]	[37.2, 38.1]	[11.1, 11.1]
Z_{100}	(m)	[39.3, 43.0]	[8.1, 8.5]	[34.4, 47.1]	[7.3, 9.7]	[43.8, 44.6]	[7.8, 7.8]
Z_{50}	(m)	[40.3, 48.6]	[6.4, 8.4]	[35.6, 52.2]	[7.0, 10.1]	[48.3, 49.4]	[7.5, 7.6]
Θ_{1500}	(°)	[24.5, 25.1]	[7.0, 7.1]	[19.4, 32.3]	[4.3, 9.9]	[26.2, 26.9]	[6.0, 6.4]
Θ_{5000}	(°)	[18.7, 19.6]	[6.2, 6.3]	[14.3, 27.2]	[4.0, 8.4]	[20.9, 21.7]	[5.3, 5.6]
Θ_{9000}	(°)	[19.4, 20.2]	[6.1, 6.1]	[14.1, 27.7]	[3.6, 8.7]	[21.7, 22.7]	[5.1, 5.4]
Θ_{16000}	(°)	[14.8, 15.3]	[5.0, 5.2]	[14.0, 21.5]	[3.5, 5.9]	[16.1, 16.9]	[4.5, 4.7]
Θ_{22000}	(°)	[20.3, 20.9]	[3.5, 3.7]	[19.5, 22.3]	[2.7, 6.1]	[20.4, 20.8]	[3.3, 3.4]
w_{1500}	(ms^{-1})	[2.29, 2.37]	[0.41, 0.43]	[2.03, 2.59]	[0.22, 0.76]	[2.37, 2.39]	[0.35, 0.36]
w_{5000}	(ms^{-1})	[2.76, 2.8]	[0.69, 0.72]	[2.26, 3.43]	[0.38, 1.11]	[2.94, 2.97]	[0.61, 0.62]
w_{9000}	(ms^{-1})	[4.10, 4.16]	[1.12, 1.12]	[3.22, 5.30]	[0.62, 1.76]	[4.44, 4.48]	[0.96, 0.98]
w_{16000}	(ms^{-1})	[3.32, 3.47]	[0.66, 0.68]	[3.02, 3.93]	[0.41, 0.97]	[3.45, 3.55]	[0.59, 0.60]
w_{22000}	(ms^{-1})	[2.21, 2.34]	[0.34, 0.39]	[2.19, 2.53]	[0.23, 0.69]	[2.31, 2.34]	[0.35, 0.36]

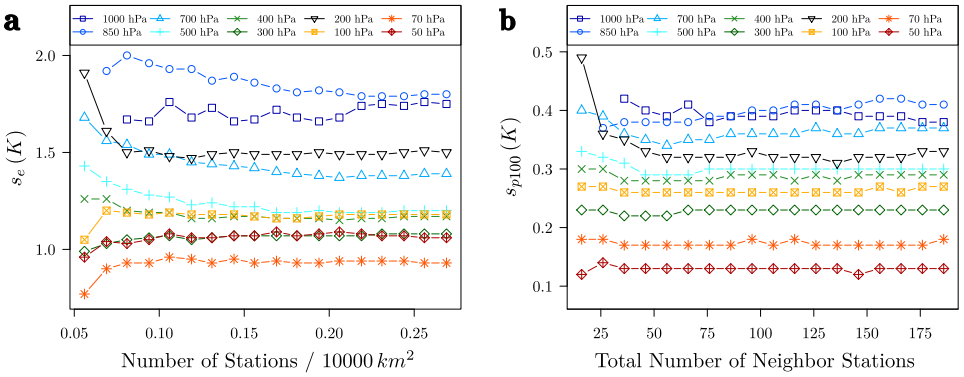

Figure A3. Temperature error estimates for stations from the contiguous United States: (a) s_e as a function of station density, (b) s_{p100} as a function of the total number of stations used to compute error estimates.

Appendix C: Parameter Choice and Sensitivity Analysis

[51] All parameters and decisions relevant for the error estimation method were tested by following a decision tree (Figure A1). For the decisions II, III, and IV, we tested the sensitivity of the method with respect to variations of the individual parameters used by our model. The optimal parameters detected in an individual decision step i were implicit for all steps j, with $j > i$. For $j \leq i$, we used $\min(n_r) = 5$, $\max(d) = 1000$, and $\Delta t = 6$ for each unknown optimal parameter.

[52] We detected significant ($\alpha = 0.05$) differences between random errors in wind speed and direction estimated from the RAW version and those estimated from the NF version (decision I; Figure A1), where the NF version contains all wind observations that were not vertically interpolated [see *Stickler et al.*, 2010]. The high vertical variability of the mean error profiles in RAW is obviously linked to levels to which a subset of observations was interpolated, as this variability is smoothed out in the NF version (Figure A2). In order to exclude interpolation errors, we decided to use NF for w and Θ and RAW for T and Z (no significant differences in the means).

[53] In decisions II to IV, we examined a number of statistics to find a combination of the parameters $\min(r_n)$, $\max(d)$, and Δt that optimizes the average performance of the error estimation method for each of the analyzed climate variables. Assuming normality of the residuals, the quality and significance of the individual linear models were estimated by the coefficient of determination (R^2), by the F test of goodness of fit to test the significance of R^2 (associated p values, p_F), and by the coefficient of variation of the model residuals (cv). The selection of the error estimation parameters was based on a weighing of these statistical indicators against the total number of error estimates (i.e., the number of candidate series c_n) and the overall mean number of reference series $[r_n]$. For $\min(r_n)$, c_n was weighted times 4 (which is equal to the number of the other parameters), as high values of $\min(r_n)$ heavily limit the number of error estimates. The results are aggregated in Table A1.

[54] Decision V tests the influence of choosing a single observation platform in favor of using observations from all platforms. Only radiosondes (T and Z) and piballs (w and Θ) were tested, as other platforms are too sparse as to allow for any error estimates to be calculated (cf. Figure 2). The differences to the original estimates of RAW (NF) were found to be not statistically significant ($\alpha = 0.05$).

[55] In decision VI, we tested the influence of the outlier removal (cf. section 3.3). This data treatment clearly reduces much of the variability of both the random error and the representativity error estimates. As it also leads to a considerable increase in the statistical significance of the linear models, we decided to use the OC version for the discussion of mean error profiles.

[56] Parallel to the estimation of the optimal parameters in decisions II to IV, we estimated the sensitivity of the error estimates with respect to variations of each individual parameter. Due to the computational complexity of the error estimation method and due to the large number of observations, it was only feasible to test the method with

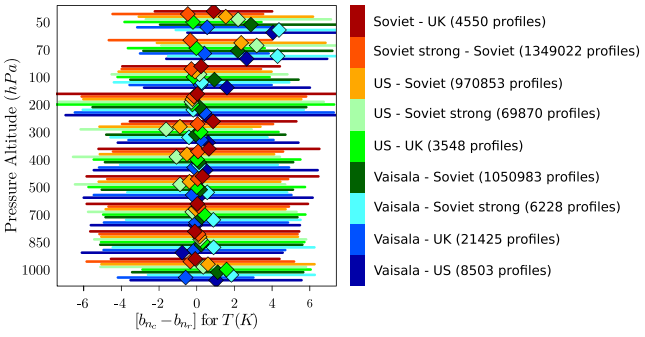

Figure A4. Profiles of mean network biases for temperatures (colored diamonds) and standard deviations of the individual difference series (bars of the same color) between all unique combinations of neighboring station networks (color key). The network "Soviet strong" contains stations from the former Soviet Union that required stronger-than-published corrections [*Grant et al.*, 2009]. Also indicated is the number of ascents available for each network pair. Values above +7 K and below −7 K are not displayed to enhance readability.

a predefined range of parameter values, as indicated in the white boxes in Figure A1. Table A2 lists the range of values of $[s_e]$ and $[s_{p100}]$ on selected altitude levels under variation of a single test parameter. It is obvious that the estimated errors are most sensitive to the choice of the neighbor search radius ($\max(d)$), while the other threshold parameters only play a marginal role.

[57] Using the previously determined model parameters, an additional experiment was performed to test the robustness of the error estimates with respect to changes in the station density. We used stations from the comparatively dense U.S. radiosonde network operated in the contiguous United States, iteratively removed 10 randomly selected stations and recomputed the error estimates. Figure A3 reveals that the mean temperature errors (averages over each available station) within the U.S. network are mostly insensitive to changes in the number of remaining neighbors (the same is true for geopotential heights; not shown). Substantial variations of the estimates can only be observed when the total number of neighbor stations drops below 25, and only for a subset of pressure levels. Note that it cannot be ruled out that the variations are influenced by the spatial distribution of the remaining stations.

[58] **Acknowledgments.** R. Wartenburger and S. Brönnimann were funded by the Swiss National Science Foundation project EVALUATE (SNF 200021-130407). A. Stickler was funded through the EU FP-7 project ERA-CLIM (265229).

References

Anderson, T. W., and D. A. Darling (1954), A test of goodness of fit, *J. Am. Stat. Assoc.*, *49*(268), 765–769, doi:10.1080/016214591954.10501232.

Bureau International des Poids et Mesures, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, and OIML (2008), *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement,* 1 ed., International Organisation for Standardization, Sèvres.

Brasefield, C. (1948), Measurement of air temperature in the presence of solar radiation, *J. Atmos. Sci.*, *5*, 147–151, doi:10.1175/1520-0469(1948)005<0147:MOATIT>2.0.CO;2.

Brönnimann, S. (2003), A historical upper air-data set for the 1939–44 period, *Int. J. Climatol.*, *23*(7), 769–791, doi:10.1002/joc.914.

Brönnimann, S., G. P. Compo, R. Spadin, R. Allan, and W. Adam (2011), Early ship-based upper-air data and comparison with the Twentieth Century Reanalysis, *Clim. Past.*, *7*(6), 265–276, doi:10.5194/cpd-6-2423-2010.

Compo, G., et al. (2011), The Twentieth Century Reanalysis Project, *Q. J. Roy. Meteor. Soc.*, *137*(654), 1–28, doi:10.1002/qj.776.

Daley, R. (1993), *Atmospheric Data Analysis*, 2, 463 pp., Cambridge Univ. Press, Cambridge.

Dee, D., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. Roy. Meteor. Soc.*, *137*(656), 553–597, doi:10.1002/qj.828.

Durre, I., R. Vose, and D. Wuertz (2006), Overview of the integrated global radiosonde archive, *J. Clim.*, *19*(1), 53–68, doi:10.1175/JCLI3594.1.

Ferguson, C. R., and G. Villarini (2012), Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States, *J. Geophys. Res.*, *117*, D05123, doi:10.1029/2011JD016988.

Francis, J. A. (2002), Validation of reanalysis upper-level winds in the arctic with independent rawinsonde data, *Geophys. Res. Lett.*, *29*(9), 29–1, doi:10.1029/2001GL014578.

Free, M., D. Seidel, J. Angell, J. Lanzante, I. Durre, and T. Peterson (2005), Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new data set of large-area anomaly time series, *J. Geophys. Res.*, *110*, D22101, doi:10.1029/2005JD006169.

Frigge, M., D. C. Hoaglin, and B. Iglewicz (1989), Some implementations of the boxplot, *Am. Stat.*, *43*(1), 50–54, doi:10.1080/00031305.1989.10475612.

Gaffen, D. (1993), *Historical Changes in Radiosonde Instruments and Practices, Instruments and Observing Methods,* vol. 50, 123 pp., World Meteorological Organization, TD-No. 541., Geneva.

Gaffen, D. (1994), Temporal inhomogeneities in radiosonde temperature records, *J. Geophys. Res.*, *99*, 3667–3667, doi:10.1029/93JD03179.

Grant, A., S. Brönnimann, T. Ewen, and A. Nagurny (2009), A new look at radiosonde data prior to 1958, *J. Clim.*, *22*(12), 3232–3247, doi:10.1175/2008JCLI2539.1.

Griesser, T., S. Brönnimann, A. Grant, T. Ewen, A. Stickler, and J. Comeaux (2010), Reconstruction of global monthly upper-level temperature and geopotential height fields back to 1880, *J. Clim.*, *23*(21), 5590–5609, doi:10.1175/2010JCLI3056.1.

Gruber, C., and L. Haimberger (2008), On the homogeneity of radiosonde wind time series, *Meteorol. Z.*, *17*(5), 631–643, doi:10.1127/0941-2948/2008/0298.

Häberli, C. (2006), *The Comprehensive Alpine Radiosonde Dataset (CALRAS)*, 4, 312 pp., Facultas, Vienna.

Haimberger, L. (2007), Homogenization of radiosonde temperature time series using innovation statistics, *J. Clim.*, *20*(7), 1377–1403, doi:10.1175/JCLI4050.1.

Haimberger, L., C. Tavolato, and S. Sperka (2008), Toward elimination of the warm bias in historic radiosonde temperature records – Some new results from a comprehensive intercomparison of upper-air data, *J. Clim.*, *21*(18), 4587–4606, doi:10.1175/2008JCLI1929.1.

Haimberger, L., C. Tavolato, and S. Sperka (2012), Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations, *J. Clim.*, *25*, 8108–8131, doi:10.1175/JCLI-D-11-00668.1.

Hoinka, K. P. (1998), Statistics of the global tropopause pressure, *Mon. Weather Rev.*, *126*(12), 3303–3325, doi:10.1175/1520-0493(1998)126<3303:SOTGTP>2.0.CO;2.

Hyndman, R. J., and Y. Fan (1996), Sample quantiles in statistical packages, *Am. Stat.*, *50*(4), 361–365, doi:10.1080/000313051996.10.473566.

Jarque, C. M., and A. K. Bera (1980), Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Econ. Lett.*, *6*(3), 255–259, doi:10.1016/0165-1765(80)90024-5.

Jasperson, W. (1982), The limiting accuracy of wind profiles obtained by tracking rising balloons, *J. Appl. Meteorol.*, *21*, 816–822, doi:10.1175/1520-0450(1982)021<0816:TLAOWP>2.0.CO;2.

Kitchen, M. (1989), Representativeness errors for radiosonde observations, *Q. J. Roy. Meteor. Soc.*, *115*(487), 673–700, doi:10.1002/qj.49711548713.

Lanzante, J., S. Klein, and D. Seidel (2003), Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison, *J. Clim.*, *16*(2), 241–262, doi:10.1175/1520-0442(2003)016<0241:THOMRT>2.0.CO;2.

McCarthy, M., H. Titchner, P. Thorne, S. Tett, L. Haimberger, and D. Parker (2008), Assessing bias and uncertainty in the hadat-adjusted radiosonde climate record, *J. Clim.*, *21*(4), 817–832, doi:10.1175/2007JCLI1733.1.

McGrath, R., T. Semmler, C. Sweeney, and S. Wang (2006), Impact of balloon drift errors in radiosonde data on climate statistics, *J. Clim.*, *19*(14), 3430–3442, doi:10.1175/JCLI3804.1.

NOAA, NASA, and USAF (1976), *U.S. Standard Atmosphere, 1976*, U.S. Government Printing Office, Washington, D.C.

OMI (Organisation Météorologique Internationale) (1951), *Comparaison Mondiale des Radiosondes. Acte Final*, vol. I, Station Centrale Suisse de Météorologie, Payerne, Switzerland.

OMM (Organisation Météorologique Mondiale) (1952), *Comparaison Mondiale des Radiosondes. World Comparison of Radiosondes. Acte Final*, vol. III, Station Centrale Suisse de Météorologie, Payerne, Switzerland.

Rossi, V. (1954), On the solar radiation error of the different radiosondes, *Eripainos Geophysica*, *4*, 195–202.

Seidel, D., C. Ao, and K. Li (2010), Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis, *J. Geophys. Res.*, *115*, D16113, doi:10.1029/2009JD013680.

Seidel, D., B. Sun, M. Pettey, and A. Reale (2011), Global radiosonde balloon drift statistics, *J. Geophys. Res.*, *116*, D07102, doi:10.1029/2010JD014891.

Sherwood, S. (2007), Simultaneous detection of climate change and observing biases in a network with incomplete sampling, *J. Clim.*, *20*(15), 4047–4062, doi:10.1175/JCLI4215.1.

Sherwood, S., C. Meyer, R. Allen, and H. Titchner (2008), Robust tropospheric warming revealed by iteratively homogenized radiosonde data, *J. Clim.*, *21*(20), 5336–5352, doi:10.1175/2008JCLI2320.1.

Stickler, A., and S. Brönnimann (2011), Significant bias of the NCEP/NCAR and Twentieth-Century reanalyses relative to pilot balloon observations over the West African monsoon region (1940–1957), *Q. J. Roy. Meteor. Soc.*, *137*(659), 1400–1416, doi:10.1002/qj.854.

Stickler, A., et al. (2010), The comprehensive historical upper-air network, *B. Am. Meteorol. Soc.*, *91*, 741–751, doi:10.1175/2009BAMS2852.1.

Thorne, P., D. Parker, S. Tett, P. Jones, M. McCarthy, H. Coleman, P. Brohan, and J. Knight (2005), Revisiting radiosonde upper air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine (2011), Tropospheric temperature trends: History of an ongoing controversy, *WIREs Clim. Change*, *2*(1), 66–88, doi:10.1002/wcc.80.

World Meteorological Organization (WMO) (2008), *Guide to Meteorological Instruments and Methods of Observation,* 8, 7 ed., World Meteorological Organization.

Xu, J., and A. Powell (2012), Uncertainty estimation of the global temperature trends for multiple radiosondes, reanalyses, and CMIP3/IPCC climate model simulations, *Theor. Appl. Climatol.*, *108*, 505–518, doi:10.1007/s00704-011-0548-z.

Zaitseva, N. A. (1993), Historical developments in radiosonde systems in the former Soviet Union., *B. Am. Meteorol. Soc.*, *74*, 1893–1900, doi:10.1175/1520-0477(1993)074<1893:HDIRSI>2.0.CO;2.